Originally published in: Online Proceedings of the 12th Mediterranean Morphology Meeting - MMM12 (2020), pp. 28-41.

Why do some lexemes combine more frequently than others? – An empirical approach to productivity in German compound formation

Katrin Hein Annelen Brunner
Leibniz Institute for the German language (IDS)
hein@ids-mannheim.de brunner@ids-mannheim.de

1. Introduction

German is well known for making extensive use of compounds. Especially nominal compounding is considered as a productive process of German word-formation (e.g. Olsen 2015; Schlücker 2012) and thus as a dominant factor in the expansion of the German lexicon. As "morphological productivity manifests itself most clearly in the appearance of complex words that never make it to the dictionary" (Booij 2012: 71), the following examples for nonce-formations from our data (section 3.1) underline the productive instantiation of the word-formation pattern in German. In other words, they illustrate its extendability "to new cases" (Booij 2012: 70):

(1) a. *Holzfrühstücksbrettchen* 'small wooden breakfast board'
b. *Fußballmacho* 'football macho'
c. *Gartenzwerg-Attitüde* 'garden gnome attitude'
d. *Immer-noch-Hippie* 'still hippie'
e. *Waschlappendieb* 'facecloth thief'
f. *Retro-Look* 'retro look'
g. *Zebrasommerwind* 'zebra summer wind'
h. *Gewohnheitsreligiosität* 'customary religiosity'
i. *Provinz-Einerlei* 'province monotony'
j. *Spätherbstnebelgrau* 'late autumn fog grey'
k. *Soundprotzerei* 'sound swank'
l. *Zwitscherwelt* 'twitter world'
m. *Standby-Spielertrainer* 'standby player coach'

Although in theory the productivity of nominal compounding is not controversial at all, in practice we can observe that certain lexemes are more frequently combined with each other than others. As a consequence, we assume with Bauer that "it is not the N+N pattern of compounding which is productive, but patterns with individual lexemes within that" (Bauer 2017: 74). But the crucial question here is: What makes a lexeme productive with respect to compound formation?

The aim of this paper is to empirically validate a factor which might influence productivity in compounding (cf. Hein and Engelberg 2018: 41-42 for a complete overview over potential factors), namely the morphological complexity of a lexeme (section 3). More precisely, we will investigate the influence of a lexeme's word-formation type on its productivity as a head-word

MMM12 Online proceedings 2020
J. Audring, N. Koutsoukos & C. Manouilidou (Eds)

in compound formation, with simultaneous consideration of frequency effects. In doing so, we draw on results from previous pilot studies (cf. Hein and Engelberg 2018).

The approach at hand is one of the first attempts to apply the concept of 'morphological productivity' (e.g. Bauer 2001, 2005), which has been predominantly applied to the domain of derivation, to compounding (section 2.1). In contrast to our previous studies, it is characterized by a stronger quantitative orientation. We present a semi-automatic approach in which quantitative measures –mainly Baayen's (1992, 2009) potential productivity measure– are applied to 100,000 nominal compound tokens which were automatically extracted from a subset of the German Reference Corpus (DeReKo, cf. Leibniz-Institut für Deutsche Sprache 2017). Needless to say, within this semi-automatic approach to compounding we are dealing with some errors in the automatic analysis (e.g. segmentation of compounds, section 3.1). Nonetheless, our investigation is worthwhile for the study of compounding as, for the first time, it allows us to empirically evaluate factors that might influence the productivity in compound formation on a large data basis.

2. Measuring compound productivity

2.1 Pitfalls

Even nowadays, "morphological productivity is one of the most contested areas in the study of word-formation" (Bauer 2001: I; cf. Bauer 2005; Plag 1999) or, as Bauer et al. (2019: 44) put it in their recent publication: "There is an extensive literature on morphological productivity, and yet the nature of productivity remains obscure." As a consequence, an answer to the question 'What does productive/unproductive mean?' is far from being trivial. In previous work (Hein and Engelberg 2018: section 2.1), we already gave an overview of the complexity of the concept, which finds expression in Rainer's (1987: 188-90) six possible readings of productivity as well as in Barðdal's (2008: 10f.) identification of 19 senses of 'productive'.

As the empirical validation of potential factors for productivity is at the heart of this paper, we do not want to repeat the whole debate about morphological productivity at this point (for an overview cf. Bauer, Beliaeva and Tarasova 2019: section 1). The core of the problem seems to be in a nutshell: Productivity is in the tension between 'availability' and 'profitability', i.e., between the theoretical possibility of new coinages and the actual exploitation of this potential (cf. Corbin 1987). This is also reflected in the question whether productivity is to be considered a qualitative or a quantitative (e.g. Scherer 2005) resp. an absolute or gradual phenomenon (e.g. Schlücker 2012: 2).

The following two general definitions of morphological productivity can serve as an adequate approximation to the concept, though even these simplifications already reflect the different readings of productivity between the poles of 'availability' and 'profitability':

i. "When we call a morphological pattern productive, we mean that this pattern can be extended to new cases, can be used to form new words." (Booij 2012: 70)
ii. "Productivity is generally defined as the extent to which some morphological process is exploited by speakers." (Bauer, Beliaeva and Tarasova 2019: 44)

In addition to its complexity, also the applicability of the notion of productivity to the domain of compounding is neither trivial nor self-evident. On the contrary, "productivity has mainly been discussed in relation to derivational morphology" (Bauer, Beliaeva and Tarasova 2019: 45), investigating the productivity of affixes (e.g. Gaeta and Ricca 2006, 2015; Scherer 2005; Hartmann 2016). The approach at hand (section 2.2; 3) is among the first to apply the concept

of morphological productivity to compounding (cf. also Tarasova 2013, 2019; Bauer, Beliaeva and Tarasova 2019; Roth 2014; Kopf 2018; Hilpert 2015).

Even if Gaeta and Ricca (2015: 848) already pointed out that a quantitative approach to productivity, as adopted in our studies, brings advantages like a "deeper understanding of the notion of productivity and the disentanglement of its diverse components", the productivity measures themselves are associated with problems which we will address in section 3.2 (cf. also Hein and Engelberg 2018 for a more detailed discussion).

2.2 Our approach

In our approach, we treat productivity as a measurable, gradual phenomenon and apply Baayen's (1992, 2009) productivity measures (for an overview cf. Hein and Engelberg 2018: section 2.2) as well as other measures which can be used to determine lexical diversity (cf. Tu, Engelberg and Weimer 2019) to a large data set.

As already mentioned above, the approach at hand (section 3) is among the first to apply the concept of morphological productivity to compounding. While Roth (2014) focuses on the competition between collocations and compounds, Kopf (2018: 61) investigates "the morphological productivity of German N+N compounding patterns from a diachronic perspective". Hilpert (2015: 143) takes a diachronic perspective as well, exploring "how the word formation process of English noun-participle compounding has developed over the past two centuries". Tarasova (2013) is interested in the influence of so called "constituent families" (e.g. all compounds containing the noun *family* in modifier or head position) on productivity: "Our knowledge of the way a noun is used in compounds is expected to be based on our previous experience with this noun as an element of a compound, and this should influence the productivity of compounds containing this noun" (Bauer, Beliaeva and Tarasova 2019: 51). As a consequence, following Tarasova (2013) resp. Bauer et al. (2019: 52) it is more suitable to "speak about the productivity of the schema defined by the presence of a particular noun in a particular position" than to judge "just the productivity of compounding as an overall pattern".

Similar to Tarasova, we focus on specific lexemes within compounds ('lexeme-based perspective') in order to determine the patterns/schemas for which productivity values are computed in a next step. But in contrast to Tarasova, we a) do not investigate the influence of so-called constituent families on productivity and b) also consider more abstract properties of the lexemes in focus. Up until now, we have always taken the head lexeme and its abstract properties as a starting point ('head-centered perspective') and then computed productivity values for groups of compounds whose head-words share some predefined properties. Via these abstract linguistic and non-linguistic properties of head lexemes, we have tried to empirically validate factors that determine productivity in compound formation from a synchronic perspective (for a detailed compilation of productivity factors cf. Hein and Engelberg 2018: section 3.1). As the head-word of a compound is crucial for its semantic and formal properties, our head-centered approach seems to be suitable.

In contrast to our recent study (section 3), our previous studies are qualitative pilot studies on the basis of manually analyzed corpus data. Their results can only be briefly summarized here:

(i) Semantic proximity between simplex words does not automatically lead to comparable productivity values with regard to the formation of compounds. For example, color words show strikingly different tendencies to occur as a head-word in compounds. Most likely, semantic proximity between the head-words seems to lead to comparable semantic patterns of compounding (e.g. 'intensifying') (cf. Hein and Engelberg 2018: section 3.2.1).

(ii) The parameter ‘frequency‘ of a simplex influences its productivity in compound formation: More frequent simplex words are more productive in compound formation than simplex words which are already unusual in isolation (cf. Hein and Engelberg 2018: section 3.2.2).

As has already been mentioned, in the meantime our project has undergone a further development: Now it is characterized by a stronger quantitative orientation. The latter results from the wish a) to distinctly increase the data basis and b) to avoid as many a priori commitments as possible (like the selection of concrete head lexemes which was necessary in our previous pilot studies).

3. Pilot study: the role of morphological complexity for producitity

In this pilot study, we focus on the potential productivity factor ‘morphological complexity‘. We pursue the question of whether the morphological complexity of a lexeme, i.e. its status as a simplex, compound or derived word, influences its productivity in compound formation.

The study of this productivity factor is relevant for several reasons: First, in the literature, there are only rare and rather speculative hints about a possible correlation between the morphological complexity of the immediate constituents and their productivity in compounding (section 3.2). Second, we are dealing with a factor that can be handled easily within our semi-automatic approach. Third, focussing on a rather formal factor like the word-formation type of the head-word can be considered a useful supplement to the more semantic-centered pilot studies we have conducted so far (section 2.2). While we evaluated in previous studies to what extent semantic similarities between constituents lead to comparable productivity values with regard to the formation of compounds, this question is now pursued for similarities concerning the morphological complexity of the constituents.

3.1 Data set

Our data is based on the “KoGra Untersuchungskorpus”, a subset (~ 7 billion tokens) of the German Reference Corpus (DeReKo, Release 2017-II, cf. Leibniz-Institut für Deutsche Sprache 2017). This corpus has a strong focus on German newspaper texts, mainly from 1990-2014, though it also contains some literary texts and even spoken language material (cf. Bubenhofer, Konopka and Schneider 2014; https://grammis.ids-mannheim.de/korpusgrammatik/6615).

The corpus was annotated with a custom word analyzer based on the Canoo Language Tools (http://www.canoonet.eu) which adds a detailed morphological analysis to each token. Using this annotation, nominal compounds were identified and extracted for our analysis.[1]

For our study, we used a random sample of 100,000 compound tokens and generated a frequency list of their base forms. Each entry comes with a morphological analysis provided by the word analyzer. The following example shows the analysis for *Feuerwehr* ‘fire department’ (literal translation: ‘fire defense’):

(cmp:N&N feuerwehr_N (feuer_N)(drv:V2N:con wehr_N (wehren_V)))

[1] Many thanks to Felix Bildhauer, who performed these extractions for us, and to our other colleagues from the project “Corpus grammar: grammatical variations in standard language and near-standard German” of the Leibniz Institute for German Language for their support!

This analysis gives us the following information:

(i) A hierarchical segmentation of the compound: *Feuerwehr* has the immediate constituents *Feuer* 'fire' and *Wehr* 'defense'. *Wehr* can be further analysed as based on the verb *wehren* 'to defend'.
(ii) Information about the word-formation type for each constituent: *Feuer* + *Wehr* is a noun-noun compound (cmp:N&N), *Wehr* is a verb-to-noun conversion (drv:V2N:con).
(iii) Word type information for each base component: *Feuer* is a noun (N), *wehren* is a verb (V).

This rich information allowed us to determine the head-word and its word-formation type for each entry. Table 1 shows a simplified excerpt from this data, which only contains the columns that will be relevant for our study.

Table 1: Simplified excerpt from our compound noun frequency list (based on a sample of 100,000 compound noun tokens)

Frequency	Lemma	Head-word	Head-word type
265	Bürgermeister	meister	simplex
232	Wochenende	ende	simplex
163	Geburtstag	tag	simplex
151	Feuerwehr	wehr	drv_con
148	Ministerpräsident	präsident	drv_suf
136	Wettbewerb	bewerb	drv_con
134	Fußball	ball	simplex
129	Arbeitsplatz	platz	simplex
129	Fahrzeug	zeug	simplex
…	…	..	…
30	Fachhochschule	hochschule	cmp

Each line represents a nominal compound type. For each type, we have the frequency information, the lemmatized form, information about its head-word and the word-formation type of the head-word. For our study, the relevant values for 'head-word type' are simplex, compound ('cmp'), conversion ('drv_con') –which is considered a type of derivation in the Canoo tagset[2]– and several other types of derivation such as suffixation ('drv_suf'), prefixation etc.

The frequency counts may seem low, but remember that this list is based on a sample that contains 100,000 tokens. This list already gives us 51,743 different compound types. 39,550 of these types have a frequency count of one, i.e. are categorized as hapax legomena. As the number of hapaxes is central to Baayen's measure of potential productivity, which we used in our study (section 3.2), we checked these words and observed that some of them are fairly established in the German language but just happen to appear only once in our sample (e.g. *Kaffeepause* 'coffee break'). This problem of "'spurious' hapaxes" (Gaeta and Ricca 2006: 68) has been noted before and occurs for any finite corpus. It is ameliorated by increasing the data basis. With the 100,000 token sample, we are already working with a collection which is considerably larger and more representative than any of our earlier studies.

[2] From a linguistic point of view, conversion can be considered either a separate type of word-formation (e.g. Fleischer and Barz 2012) or a sub-type of derivation (e.g. Booij 2012).

As the information about head-word and head-word type is based on the automatic analysis of the Canoo tool, one may wonder how reliable this information is. To judge its quality, a manual evaluation was performed.[3] We checked for the following issues:

(i) Is the *automatic segmentation* into immediate constituents correct? (sample: 500 compound types) → ca. 87,6 % unproblematic cases
(ii) Is the *head-word* detected (and lemmatized) correctly? (sample: 800 head-words) → ca. 88 % unproblematic cases
(iii) Is the *word-formation type* assigned to the head-word correct (simplex vs. compound vs. derived word)? (sample: 800 head-words) → 88,63 % unproblematic cases

Overall, we conclude that the Canoo analysis is of good quality and we feel confident to use this data for our study. Although there are errors, their percentage is presumably low enough to not distort the quantitative results for the whole sample, and the benefits of working with a truly large sample of compounds outweighs the unavoidable inaccuracies that come with automatic analysis.

3.2 Basic assumptions & approach

As already mentioned above, our recent study is about the relation between morphological complexity and productivity. More precisely, we are interested in the influence of a lexeme's word-formation type (simplex vs. derived word vs. compound word) on its productivity as a head-word in compound formation.

It can be considered a consensus for German that the immediate constituents of a compound can be both simplex words and complex words, i.e. the products of word-formation, and that all kinds of complex words can serve as head-word or modifier within a compound (cf. Ortner and Ortner 1984: 116f.). However, it does not seem plausible to assume that there are no differences in productivity among these different types of compound constituents. As a starting point, we then assume with Fleischer and Barz (2012: 81) that the relation between the morphological complexity of the roots and their productivity with regard to compound formation is different than in derivation. The following hypothesis is adopted: "While the affinity to derivation decreases clearly with an increasing complexity of the basis, this cannot be said in the same way for composition" (Fleischer and Barz 2012: 81, our translation).

According to Fleischer and Barz (2012), the situation in derivation is as follows: While one can find many derived words for non-complex bases like *krank* 'sick' (e.g. *erkranken* 'to sicken'*, Krankheit* 'sickness'*, krankhaft* 'pathological'), there is often only one derived word for more complex stems like *kränklich* 'sickly' (e.g. *Kränklichkeit* 'sickliness'). Those rather vague, but interesting hints have initiated this study. As, to our knowledge, there are no comparable assumptions in the literature with regard to the relation between a lexeme's morphological complexity and its productivity as a head-word in compound formation, this issue seems to be highly relevant for research on German word-formation.[4] A semi-automatic approach on the basis of large data seems to be particularly suitable to illuminate this issue. Thus, the aim of the study at hand is to specify Fleischer and Barz' vague, non-corpus-based hints on a large data basis. Please note that Fleischer and Barz do not focus on the question of

[3] Many thanks to Lena Stutz for her work!
[4] Concerning the morphological complexity of the non-head, one can assume a restriction for the case of German A+N compounds: "[…] with regard to morphological structure, only monomorphemic adjectives are allowed" (Schlücker and Hüning 2009: 212f.).

productivity in the sense of our study. Nonetheless, their key concept 'composition activity' („Kompositionsaktivität") (cf. Fleischer and Barz 2012: 81f., 133f., 152f.) seems to be roughly transferable to the question of morphological productivity.

As far as quantitative measures are concerned, we mainly focus on potential productivity, one of several productivity measures defined by Baayen (1992, 1993, 2001, 2009). Within a group of compounds, this measure is calculated as $\frac{number\ of\ hapax\ legomena}{total\ number\ of\ tokens}$. It is intended to measure the degree of saturation of a word-formation pattern (Baayen 2009: 902). This is of course just one aspect of the complex concept of productivity as outlined in section 2.1. The idea is that hapaxes, words which occur only once in the data set, are a good approximation for newly coined words. Though this measure has been widely adopted, it has been criticized for its dependency on the number of tokens (e.g. Roth 2014; Gaeta and Ricca 2006: 61) as well as for its reliance on hapax counts: "[…] it is essential to bear in mind that the number of hapaxes is only an indirect indicator of the rate of expansion of a morphological category, since the number of hapaxes is not a direct reflection of the number of neologism coined by a given morphological process" (Bauer, Beliaeva and Tarasova 2019: 49).

Following suggestions from Tu, Engelberg and Weimer (2019), we also experimented with a second measurement, which can be used to capture the concept of lexical diversity: entropy. Entropy is a measure from information theory (cf. Shannon 1948), which has already been adopted for linguistic studies (e.g. Gibson et al. 2019). It can be interpreted as the uncertainty when drawing a random element from a group, such as a single token from a collection of compound nouns. Entropy is low if there are few very frequent types: In this case, one can be fairly certain which compound token one will get. In the extreme case that there is only one compound type, entropy would be zero. On the other hand, entropy is high if there are many different, uniformly distributed types in the group. Thus, entropy can measure how diverse a group is. This is not necessarily the same as how productive a word-formation pattern is, but it may be related to productivity.[5]

We calculated entropy scores in addition to potential productivity scores at all steps of our analysis. In this study, our focus –especially with respect to linguistic interpretation– will remain on the potential productivity measure, but we will address whether entropy suggests similar or divergent rankings of our groups. If we consider entropy as a measure for lexical diversity and thus related to productivity, similar rankings would strengthen the quantitative results.

To evaluate the potential influence of the factor 'morphological complexity of the head' on its compound productivity, we proceed as follows: In a first step, we define three main groups, according to abstract properties of the head lexemes:

(i) 'simplex': compounds whose head is a simplex (e.g. *Macho* 'macho' in *Fußball-Macho* 'football macho')
(ii) 'cmp': compounds whose head is a compound itself (e.g. *Spielplatz* 'playground' in *Abenteuerspielplatz* 'adventure playground')
(iii) 'drv': compounds whose head is a derived word (e.g. *Regierung* 'government' in *Bundesregierung* 'federal government')

[5] For a group such as 'noun compounds with a simplex head-noun' entropy is calculated with the following formula: $Entropy = -\sum_{i=1}^{number\ of\ types} \frac{\text{number of tokens of type}_i}{\text{total number of tokens}} * \log_2 \left(\frac{\text{number of tokens of type}_i}{\text{total number of tokens}}\right)$

We now want to measure potential productivity and entropy for each of these groups and compare them.

However, when sorting the compounds from our list into the three groups, we arrive at groups with markedly different sizes. Especially the group 'cmp' –compounds with a head that is a compound itself– is much smaller than the two other groups (cf. Table 2). This is no surprise, as this group requires compounds with at least three components and at least two composition processes ('recursion'), which is said to be less common for German than binary compounds which do not show recursion (Ortner et al. 1991: 9).

Table 2: Type and token counts of the compounds per head-word type in our 100,000 token sample

head-word type	compound types	compound tokens
cmp	3,314	4,446
drv	22,204	39,285
simplex	25,305	54,951

Though this frequency distribution among the groups is an interesting observation in itself, it poses a methodological problem which is well-known in the literature. The measurement of potential productivity is dependent on the token size of the observed group, so groups with different tokens sizes cannot be compared in a meaningful way: "The ratio h/N [hapax legomena/tokens] does not seem to give meaningful results if, in a given corpus, one compares the results obtained for affixes with very different token frequencies" (Gaeta and Ricca 2006: 62). This dependency on token size is true for entropy as well (cf. Tweedie and Baayen 1998). This problem is especially grave for the 'cmp'-group, which is smaller than the other two groups by a factor of 10.

Our solution to this problem is repeated sampling (cf. Tu, Engelberg and Weimer 2019): From each group, the same fixed number of compound tokens is picked randomly. We decided to use a sample size of 2,000, which is large enough to give a good representation and still allows repeated sampling without too much overlap for our smallest group (4,446 tokens). Now we have groups of the same size. For these, we calculate our measures, potential productivity and entropy, and save the results. We execute this sampling process 1,000 times, picking new random samples and calculating the scores for each iteration. This repetition helps to generate more reliable results – if we only sample once, it is possible that we pick a skewed sample that would lead us to wrong conclusions, but after 1,000 picks, we can be fairly certain to have a good representation of our data.

After sampling, we have 1,000 values for potential productivity and for entropy for each group. To visualize those, we use kernel density plots (cf. Cox 2007), a type of plot that is helpful to show the distribution of a value. We will explain in more detail how to read those plots below (section 3.3), when we show our results. As mentioned above, we will focus on potential productivity in this study. Entropy scores will therefore not be reported in full, but only be addressed briefly in comparison to potential productivity scores.

3.3 Results: overview and linguistic interpretation

We will now present the results of our quantitative analysis and interpret the findings from a linguistic perspective. In this context, we will consider structural assumptions from the literature in order to evaluate if the quantitative observations are to be expected, and also address concerns about factors which might distort the quantitative analyses.

3.3.1 Simplex head vs. complex head

In the first part of our study, we compared compounds with a morphologically complex head-word on the one hand and compounds with a morphologically simple head-word on the other hand. Compounds and derived words in head-position count as morphologically complex, simplex words count as not complex. From a technical perspective, this means we combine the compound tokens in groups 'cmp' and 'drv' and compare those to the compound tokens in 'simplex'.

Figure 1: Potential productivity (sample size: 2,000; complex vs. simplex)

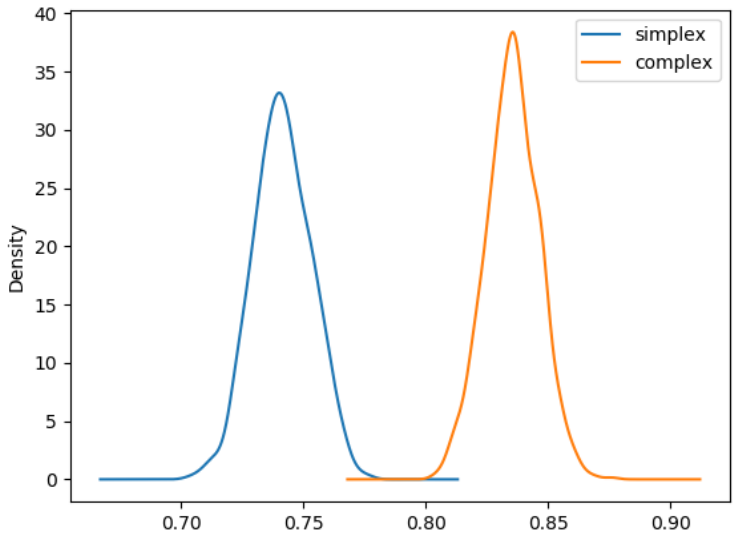


Figure 1 shows the results of the quantitative analysis. The density plot displays how often a certain value occurs over the 1,000 sampling runs. The x-axis shows the actual values for potential productivity. The further to the right a distribution is located, the higher are its values. In our case, the orange curve for complex head-words is located clearly to the right of the blue curve for simplex head-words, showing consistently higher values for potential productivity. The y-axis represents the density of the distribution. High, steep peaks indicate that similar values occur very often in the 1,000 sampling runs. Low and wide curves would indicate a broad spread of different values. This is not the case in this evaluation – both groups behave very similarly over repeated sampling runs.

Our quantitative analysis thus indicates that morphologically complex words are more productive as head-words in compounds than simplex words. This finding might be expected if productivity –as in our case– is based on the percentage of hapaxes: While compounds with simplex bases are often highly frequent lexicalized words (e.g. *Arbeitsplatz* 'work place'), one can assume a higher number of hapaxes in the group of compounds with morphologically complex head-word (e.g. *Kindermaskenball* 'children's masked ball', *Ballfänger* 'ball catcher') which are rarer and more diverse. As already mentioned above, the hapax centricity of the measure 'potential productivity' might have an influence on the results.[6]

How are the results in Figure 1 to be judged from a linguistic perspective? Our findings are in line with principal assumptions in the literature, namely that the immediate constituents of a

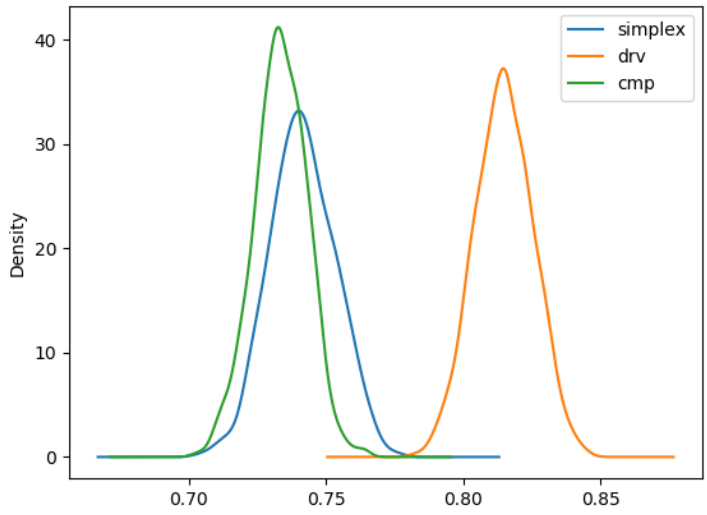
[6] It should be noted though, that we get a very similar picture if we use the entropy measure: The group of compounds with complex head-words has clearly a higher entropy than the group with simplex head-words. It can thus be considered a more diverse and less predictable group.

compound can be formed by both simplex words and word-formations (section 3.2) and that composition in German allows for recursion (Schlücker 2012: 8). Nonetheless, the higher productivity of morphologically complex head-words in comparison to simple head-words is not necessarily expected, considering that complex constituents are said to be instantiated more often in the modifier position than in the head position of a compound (cf. Ortner and Ortner 1984: 116f.), and that –as a consequence– left-branching compounds are said to be more common/frequent than right-branching-compounds (cf. Schlücker 2012: 8; Libben 2016: 19). That is why, in the future, it might be important to additionally consider also the morphological complexity of the non-heads.

3.3.2 Compound vs. derived word vs. simplex as head

In a second analysis, we distinguished between compounds with a simplex, with a compound and with a derived word in head-position, i.e. we used the groups described in section 3.2. Conversion was subsumed under derivation in this case (section 3.1). The results are as follows: We found that derived words have the highest productivity values. Compounds and simplex words, however, show very comparable values as head-words in compounds and are both less productive than derived words.

Figure 2: Potential productivity (sample size: 2,000; compound vs. derived word vs. simplex)



The density plot in Figure 2 displays the quantitative results. The curve for derived words is located furthest to the right, indicating the highest values of potential productivity. The curves for simplex and for compound words are further to the left and show a clear overlap. This overlap means that repeated sampling runs often returned the same values of potential productivity for both of these groups. In contrast to the derived words, they do not seem distinguishable with respect to their potential productivity.

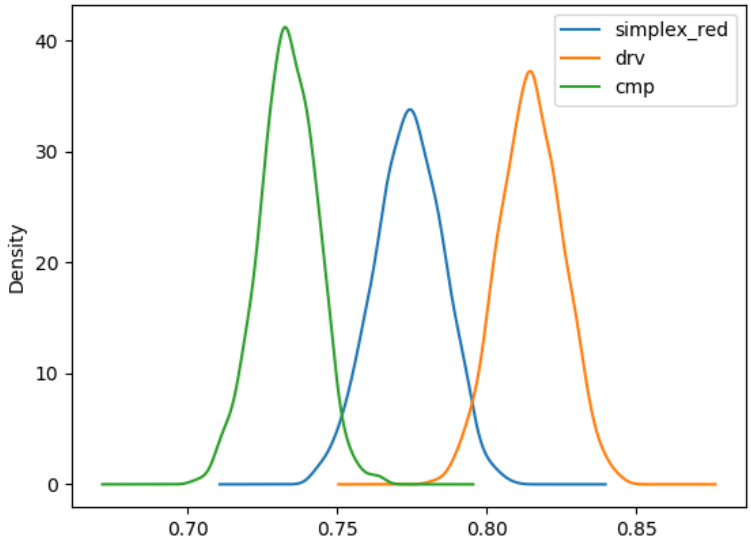
In simplified terms –as described above– the density plot shows that derived words are the most productive, while compounds and simplexes are quite similar with regard to the formation of compounds. Are these findings in line with principal structural assumptions from the word-formation literature? Ortner et al. (1991: 9) found that 80-90 % of German nominal compounds are bipartite in the sense that they do not show recursion, e.g. *Erdöl* ‘mineral oil’. The high productivity of derived head-words (e.g. *Regierung* in *Bundesregierung* 'federal government') is principally in line with this assumption. It might be that synthetic compounds (e.g. *Familiengründung* ‘family foundation’) play an important role for the high productivity of derived head-words. However, it was not necessarily expected that compounds (e.g. *Spielplatz*

'playground' in *Abenteuerspielplatz* 'adventure playground') and simplex words (e.g. *Macho* 'macho' in *Fußball-Macho* 'football-macho') display such similar productivity values. If indeed 80–90 % of German compounds do not show recursion, i.e. do not show an immediate constituent which itself has the status of a compound, one would expect that simplexes are much more productive than compounds in head position.

As our findings are –to some extent– surprising from a theoretical point of view, we searched for the potential influence of a quantitative distortion of the results. First, we checked if the higher productivity values of derived head-words are due to their generally higher occurrence as stand-alone words. The underlying assumption here is that lexemes that are more frequent in isolation are also more productive with regard to compound formation (section 2.2) (cf. Hein and Engelberg 2018: 46-49). However, the frequencies of our head-words (as stand-alone words) in the same corpus that was used to extract the compounds show that a quantitative distortion is not probable: Derived words are on average less frequent than simplex words. A question that remains open is whether the high productivity of derived head-words can be attributed to a high proportion of synthetic compounds in the data.

Linguistically even more unexpected than the clear dominance of derived head-words is the finding that compound and simplex head-words are quite similar. We checked our compound noun collection for possible distortions and noticed that it was dominated by six highly frequent and highly lexicalized words, which were categorized as having a simplex head: the days of the week (like *Montag* 'Monday').[7] The non-head of these words is almost always a cranberry morpheme. They can be considered a special case due to their high degree of lexicalization. It is also likely that the composition of our corpus, which contained a lot of newspaper data, led to an overrepresentation of such nouns. Thus, we removed the six days of the week and repeated our analysis. And indeed, this modified evaluation leads to different results (cf. Figure 3).

Figure 3: Potential productivity (without days of the week)
(sample size: 2,000; compound vs. drived word vs. simplex)



[7] The words *Montag* 'Monday', *Dienstag* 'Tuesday', *Donnerstag* 'Thursday', *Freitag* 'Friday', *Samstag* 'Saturday' and *Sonntag* 'Sunday' are the six most frequent types in our data. *Mittwoch* 'Wednesday' is missing, as it is not analyzed as a compound by Canoo.

As the data sets for derived head-words and compound head-words were not modified, the position of these groups remain the same in the new density plot. However, the simplex group, i.e. compounds with a simplex head, shifts to the right and no longer overlaps significantly with the compound word group. This means the potential productivity for the simplex group without the days of the week is indeed consistently higher than for the compound word group, although still lower than for the derived word group.[8]

According to this modified evaluation, simplexes are more productive as head-word in compounds than compounds. From a linguistic perspective, this is a more convincing finding (in comparison to Figure 2), considering that according to Ortner et al. (1991: 9) 80-90 % of German compounds do not show recursion.[9] In any case, possible distortions of the results by the occurrence of highly frequent compounds will have to be addressed in greater detail in the future.

4. Conclusion & Outlook

Even if the results give still rise to questions and require further investigation, our study indicates clearly that 'morphological complexity' is a factor that influences a lexeme's productivity in compound formation: We get different productivity values for groups of compounds with different morphological types of head-words.

To guard against errors, we did three more things: First, we pulled a second sample of 100,000 tokens from our corpus and repeated the analysis presented here. All results remained the same. This means, that it is very unlikely that our results are due to random distortions in our original 100,000 token sample. Secondly, we created random artificial groups of compound nouns and performed the sampling and plotting on those. The results were largely overlapping curves, rather than the clear distinctions visible in the plots above for the "real" groups 'complex vs. non-complex head-word' resp. 'simplex vs. compound vs. derived word'. Thus, we conclude that the distinction between different types of head-words is indeed a meaningful factor related to potential productivity. Thirdly, we observed that our compound noun collection also contained some proper names (e.g. *Karl-Ernst, Ostdeutschland* 'East Germany', *Blücherstraße* 'Blücher Street'). Worried that these might distort our findings, we removed all compound types that were identified as proper names by an automatic morpho-syntactic tagger[10] and again repeated the analysis. Still, our results remained stable.

---

[8] For the entropy scores of these three groups one can observe a similar, but even stronger effect of the days of the week: When calculating the scores with the data that includes the days of the week, the simplex group not only shows much overlap with the compound group but even tends to score lower on entropy – i.e. it appears that compounds with a simplex head are the most predictable. When removing the days of the week, the curve for the simplex group shifts sharply to the right, now even somewhat overlapping with the derivation group. From a quantitative viewpoint, this is explainable: If a group contains types with a very high frequency, there is a higher likelihood that one of these types is found in a random draw. The group is more predictable and thus has a lower entropy. In conclusion, when controlling for the days of the week, entropy again gives the same ranking of our groups as potential productivity.

[9] The observation that the group of compounds with simplex head tends to contain very frequent and lexicalized words (like the days of the week) together with the observation that compounds with compound heads are less frequent (according to Ortner et al. and also according to the frequency counts in our sample) could also lead to a different expectation: Being less frequent and less lexicalized, the compound group should contain more hapaxes, therefore scoring higher in potential productivity than the simplex group. However, when repeated sampling is employed to normalize the number of tokens for each group, this intuition is not supported.

[10] We used the TreeTagger (cf. Schmid 1995). According to a manual evaluation we did on 500 compound types, this tagger has a high precision (0.9), but not a very good recall (0.39). This means that several proper names still

The investigation of morphological complexity is also interesting beyond the question of morphological productivity. Among other things, it contributes to a corpus-based exploration of compound formation in German, as it requires a detailed description of abstract formal patterns that can be observed in compounds.

Even though our investigation of morphological complexity as a potential factor for compound productivity with the help of semi-automatic methods on the basis of a large data set was quite fruitful, some critical issues should not go unmentioned: First, it could be called into question that productivity factors can be treated as isolated properties of lexemes (without considering further properties of the pattern in which the lexemes occur). Second, in the semi-automatic study at hand we have to accept errors in the automatic analysis, and we are also dealing with general mathematical effects and issues of frequency which require further research. Anyway, the question why some lexemes combine more frequently than others will pursued further in the near future.

In the long run we aim at …

(i) … repeating the same procedure for the non-heads (and investigating the correlation between abstract properties of the non-head and the head at the same time);
(ii) … conducting comparative studies between different types of corpora, such as web corpora vs. traditional written language corpora;
(iii) … expanding our approach towards other quantitative methods like machine learning that allow us to consider several productivity factors at the same time. After all, it can be assumed that "productivity is influenced by multiple factors simultaneously" (Bauer, Beliaeva and Tarasova 2019: 76).

References

Baayen, H. R. 1992. Quantitative aspects of morphological productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology (1991)*. Dordrecht: Kluwer Academic Publishers, 109-149.

Baayen, H. R. 1993. On frequency, transparency and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology (1992)*. Dordrecht: Kluwer Academic Publishers, 181-208.

Baayen, H. R. 2001. *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, H. R. 2009. Corpus linguistics in morphology: morphological productivity. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An international handbook*. Berlin: De Gruyter Mouton, 900-919.

Barðdal, J. 2008. *Productivity. Evidence from case and argument structure in Icelandic*. Amsterdam/Philadelphia: Benjamins.

Bauer, L. 2001. *Morphological productivity*. Cambridge: Cambridge University Press.

Bauer, L. 2005. Productivity: Theories. In P. Štekauer & R. Lieber (Eds.), *Handbook of word-formation.* Dordrecht: Springer, 315-334.

Bauer, L. 2017. *Compounds and compounding*. Cambridge: Cambridge University Press.

Bauer, L., Beliaeva N. & E. Tarasova. 2019. Recalibrating Productivity: Factors Involved. *Zeitschrift für Wortbildung / Journal of Word Formation* 3 (1): 44-80. https://doi.org/10.3726/zwjw.2019.01.02.

Booij, G. 2012. *The Grammar of Words. An introduction to linguistic morphology*. Third edition. Oxford: Oxford University Press.

Bubenhofer, N., Konopka, M. & R. Schneider (Eds.). 2014. *Präliminarien einer Korpusgrammatik.* Tübingen: Narr.

Corbin, D. 1987. *Morphologie dérivationnelle et structuration du lexique.* Volume 1. Tübingen: Niemeyer.

Cox, N. J. 2007. Kernel estimation as a basic tool for geomorphological data analysis. *Earth Surface Processes and Landforms* 32 (12): 1902-1912. https://doi.org/10.1002/esp.1518.

---

remained in the sample. Still, if proper names had a very strong effect one would expect the scores to shift even when removing them only partially.

Fleischer, W. & I. Barz. 2012. *Wortbildung der deutschen Gegenwartssprache*. Fourth edition. Berlin/Boston: De Gruyter.

Gaeta, L. & D. Ricca. 2006. Productivity in Italian word formation: a variable-corpus approach. *Linguistics* 44: 57-89.

Gaeta, L. & D. Ricca. 2015. Productivity. In P. O. Müller, I. Ohnheiser, S. Olsen & F. Rainer (Eds.), *Word-formation. An International Handbook of the Languages of Europe. Volume 2, IV: Rules and restrictions in word-formation I: General aspects*. Berlin/Boston: De Gruyter Mouton, 842-858.

Gibson, E., Futrell, R., Piantadosi S. P., Dautriche, I., Mahowald, K., Bergen, L. & R. Levy. 2019. How Efficiency Shapes Human Language. *Trends in cognitive sciences* 23 (5): 389-407.

Hartmann, S. 2016. *Wortbildungswandel: Eine diachrone Studie zu deutschen Nominalisierungsmustern*. Berlin: De Gruyter.

Hein, K. & S. Engelberg. 2018. Morphological variation: the case of productivity in German compound formation. In Koutsoukos, N., Audring, J. & F. Masini (Eds.), *Proceedings of the Eleventh Mediterranean Morphology Meeting*. Patras: Pasithee, 36-50. https://doi.org/10.26220/mmm.2871.

Hilpert, M. 2015. From hand-carved to computer-based: Noun-participle compounding and the upward-strengthening hypothesis. *Cognitive Linguistics* 26 (1): 1-36.

Kopf, K. 2018. The role of syntax in the productivity of German N+N compounds. A diachronic corpus study. *Zeitschrift für Wortbildung / Journal of word formation* 2 (1): 61-91. https://doi.org/10.3726/zwjw.2018.01.03.

Leibniz-Institut für Deutsche Sprache. 2017. *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2017-II* (Release vom 01.10.2017). Mannheim: Leibniz-Institut für Deutsche Sprache. www.ids-mannheim.de/DeReKo.

Libben, G. 2016. Why Study Compound Processing? An overview of the issues. In G. Libben & G. Jarema (Eds.), *The representation and processing of compound words*. Oxford/New York: Oxford University Press, 1-22.

Olsen, S. 2015. Composition. In P. O. Müller, I. Ohnheiser, S. Olsen & F. Rainer (Eds.), *Word-formation. An International Handbook of the Languages of Europe. Volume 1, II: Units and processes in word-formation I: General aspects*. Berlin/Boston: De Gruyter Mouton, 364-386.

Ortner, H. & L. Ortner. 1984. *Zur Theorie und Praxis der Kompositaforschung*. Tübingen: Narr.

Ortner, L., Müller-Bollhagen, E., Ortner, H., Wellmann, H., Pümpel-Mader, M. & H. Gärtner. 1991. *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache . Volume 4: Substantivkomposita*. Düsseldorf: Schwann.

Plag, I. 1999. *Morphological productivity: structural constraints in English derivation*. Berlin: De Gruyter Mouton.

Rainer, F. 1987. Produktivitätsbegriffe in der Wortbildungstheorie. In: W. Dietrich & H.-M. Gauger (Eds.), *Grammatik und Wortbildung romanischer Sprachen: Beiträge zum Deutschen Romanistentag in Siegen, 30.9.-3.10.1985.* Tübingen: Narr, 187-202.

Roth, T. 2014. *Wortverbindungen und Verbindungen von Wörtern. Lexikografische und distributionelle Aspekte kombinatorischer Begriffsbildung zwischen Syntax und Morphologie*. Tübingen: Francke.

Scherer, C. 2005. *Wortbildungswandel und Produktivität. Eine empirische Studie zur nominalen -er-Derivation im Deutschen*. Tübingen: Niemeyer.

Schlücker, B. & M. Hüning. 2009. Compounds and phrases. A functional comparison between German A+N compounds and corresponding phrases. *Italian Journal of Linguistics / Rivista di Linguistica* 21 (1): 209-234.

Schlücker, Barbara. 2012. Die deutsche Kompositionsfreudigkeit. Übersicht und Einführung. In L. Gaeta & B. Schlücker (Eds.), *Das Deutsche als kompositionsfreudige Sprache. Strukturelle Eigenschaften und systembezogene Aspekte.* Berlin: De Gruyter, 1-25.

Schmid, H. 1995. Improvements in Part-of-Speech Tagging with an Application To German. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, 47-50.

Shannon, C. E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27: 379-423.

Tarasova, E. 2013. *Some new insights into the semantics of English N+N compounds*. Ph D. thesis. Victoria University of Wellington: unpublished manuscript. https://core.ac.uk/download/pdf/41338059.pdf.

Tarasova, E. 2019. Productivity of form and productivity of meaning in N+N compounds. *SKASE Journal of Theoretical Linguistics* 16 (1): 49–69. http://www.skase.sk/Volumes/JTL39/pdf_doc/04.pdf.

Tu, N. D. T., Engelberg, S. & L. Weimer. 2019. „Was für Enthüllungen!“, heulte die wohlgekleidete respektable Menge. Eine korpuslinguistische Untersuchung zur lexikalischen Vielfalt von Redeeinleitern. *Redewiedergabe. Linguistische Berichte – Sonderhefte* 27: 13-53.

Tweedie, F. J. & H. R. Baayen. 1998. How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities* 32 (5): 323-52.